PSO 알고리즘 기반 OpenMind 시스템 개선 및 효과 검증

원태연*, 양승윤**, 김정명**, 원일용**, 김현정***
*서울호서전문학교 컴퓨터공학과
**서울호서전문학교 사이버해킹보안과
***건국대학교 상허교양대학
e-mail : teadone@naver.com
supersatori@naver.com
eyy684@naver.com
clccclcc@shoso.ac.kr
nygirl@konkuk.ac.kr

Improvement and effect verification OpenMind system based on PSO algorithm

Tae-Yeon Won*, Seung-Yun Yang**, Jung-Myoung Kim**, Ill-Young Weon**, Hyun-Jung Kim***
*Dept. of Computer Science, Seoul Hoseo College
**Dept. of Cyber Hacking Security, Seoul Hoseo College
***Sang-Huh College, Kon-Kuk University

요 약

여러 분야에서 각광받는 딥러닝은 학습시간이 오래 걸리고, 고가의 장비들이 요구된다. 이러한 이유로 저사양 머신들을 이용한 분산 러닝 시스템들이 연구되기 시작했다. 본 논문은 " PSO 알고리즘을 이용한 분산 딥러닝 시스템" 을 개선했고, 그 결과 개선한 시스템의 머신 개수가 1 대 일 때 정확도가 92.8%까지 향상되었고, 머신 개수가 10 대 일 때 정확도가 93.4%까지 향상되었다. 이를 기반으로 저사양의 머신들을 결합한 분산 러닝 시스템이 고가의 장비를 사용하지 않고도 좋은 결과를 얻을 수 있다는 것을 확인했다.

1. 서론

현재 딥러닝은 4 차 산업혁명을 맞이하면서 각광받는 분야 중 하나이다[1]. 특히 딥러닝은 자율주행 자동차나 번역기 등 일상생활에서 응용되고 많은 도움이 되고 있다[2][3]. 그러나 많은 사전 데이터가 필요하며, 그 데이터를 학습시키는 것에 오랜 시간이 소요된다는 문제가 있다. 이를 해결하기 위해 병렬 연산에 최적화된 고가의 GPU 를 사용한다[4].

그 결과 GPU 를 사용함에 있어서 단일 CPU 로는 많은 시간이 걸릴 수 있는 학습도 훨씬 단시간에 학습이 가능해졌다. 그러나 NVIDIA Titan XP 와 같은 고가의 GPU 를 사용하는 사람도 적을 뿐만 아니라 대다수는 CPU 만 있는 시스템을 사용한다.

단일 CPU 시스템만을 사용한다면 학습은 굉장히 오랜 시간이 소요되며, 데이터의 형태에 따라 모델의 효율성도 변하기 때문에 데이터에 맞는 모델을 결정하는 것까지 굉장히 오랜 시간과 시행착오를 거친다. 이를 해결하기 위해 다양한 분산처리 방법들이 연구되고 있다. 여러가지 방법들 중 " PSO 알고리즘을 이용한 분산 딥러닝 시스템" 이라는 연구가 있다[5][6].

기존의 " PSO 알고리즘을 이용한 분산 딥러닝 시스템" 논문은 딥러닝이 많은 시간을 소모하는 학습단계에서 고가의 하드웨어가 아닌 저 사양의 장비를 여러 대 결합한 분산 러닝 시스템에 대한 성능 실험이다.

그러나 이 논문에서는 UCI 저장소(UCI Machine Learning Repository)에서 제공했던 ' 피마족 인디언의 당뇨병 발생 데이터 셋' 을 사용하여 정확도를 실험했고 그에 따른 성능은 매우 높게 도출됐다. 그 이유는 한정적인 데이터 셋을 사용하여 학습과 테스트를 진행한 결과 과적합(Overfitting)이 발생했기 때문이다[7].

또한 작은 데이터 셋의 크기와 모델이 단순했기 때문에 통신 비용이 적었으나 데이터 셋과 모델의 크기가 커진다면 서버와의 통신비용이 늘어나는 문제점이 있다. 따라서 본 논문에서는 이러한 문제점을 개선하고 단일시스템부터 분산시스템의 개수가 시간대비 성능에 있어서 얼마나 영향을 주는지를 검증하고 한다.

2. 관련 연구

2.1 딥러닝

딥러닝은 기계 학습의 부분 집합이고 핵심은 분류를 통한 예측이다. 딥러닝은 인공신경망의 원리를 이용해 인간의 두뇌 연결성을 모방하여 데이터를 분류하며 상관관계를 찾아낸다. 딥러닝은 비지도학습, 지도학습 이렇게 두가지 방식이 있다.

지도학습은 레이블이 주어진 상태에서 학습한다는 뜻으로 훈련용 데이터로부터 하나의 함수를 추론하고 특징변수와 목표변수 사이의 관계를 학습한다. 따라서 지도학습은 명확한 input 과 output 이 존재한다. 이러한 지도학습에 분류와 예측이 있다. 비지도학습은 데이터에 대한 레이블 이

주어지지 않은 상태에서 컴퓨터를 학습시키는 방법론이다. 즉, 데이터 형태로 학습을 진행하는 방법이다.

2.2 딥러닝 분산처리

딥러닝의 학습 시간을 줄이거나, 학습 효율을 높이기 위하여 분산 처리를 이용한다.

딥러닝 분산은 크게 모델 분산과 데이터 분산으로 나누어 진다. 모델 분산은 큰 사이즈의 모델을 여러 노드에 나누어서 보관하고 업데이트하는 방법이고 데이터 분산은 데이터를 분할하여 모델의 학습을 병렬로 진행하는 방법이다.

데이터 분산을 위해서는 모델의 복제가 이루어 지는데 이러한 복제된 모델들의 동기화 방법에 따라 동기적 동기화와 비동기적 동기화로 나누어진다. 또한 분산의 구조는 분산을 위한 노드들의 역할이나 커뮤니케이션 방식에 따라 파라미터 서버 구조와 집단 통신 구조로 나뉠 수 있다[4].

2.3 PSO 알고리즘

PSO 알고리즘은 1995 년에 Kennedy 와 Eberhart 에 의해 소개되었으며, 새 떼와 물고기 떼와 같은 생체군집의 사회적 행동양식을 모방하여 개발된 알고리즘이다.

PSO 는 휴리스틱 기법으로 컴퓨터 과학 분야에서 사용하는 일종의 경험적 전역 최적화 기법이다. 전역 최적화 기법은 시간이 오래 걸리더라도 전체 탐색영역에서 가장 좋은 해를 찾는 것을 목표로 하며, 일부 탐색영역 내에서 가장 좋은 해를 찾는 것을 목포료 하는 지역최적화 기법과는 대조된다.

PSO 에서 각 입자(particle)는 다차원 탐색공간을 옮겨다니며 다른 입자들과 정보를 교환한다. 그리고 자신과 이웃한 입자의 경험 정보를 이용하고 아래와 같은 식을 적용하여 최적의 해로 이동해 간다[11][13].

(1) v=v+ c1*rand*(pbest-present)+ c2*rand*(gbest-present)

(2) present=present+ v

위의 식(1)에서는 v 는 객체의 속도 벡터를 의미하고, c1, c2 는 학습 요소를 의미하며, rand 는 0~1 사이의 임의 숫자를 의미한다. 그리고 pbest 는 객체가 지금까지의 탐색 중 발견한 최량해의 위치 벡터를 의미하고, gbest 는 전체 객체가 지금까지의 탐색 중 발견한 최량해의 위치 벡터를 의미한다. 위의 식(2)에서 present 는 객체의 위치 벡터를 의미한다.

2.4 OpenMind

OpenMind 는 여러 시스템들로 구성된 분산 시스템 환경과 중앙에 저장소를 둔 분산 딥러닝 시스템으로 제안됐다. 이 시스템은 저사양의 하드웨어만으로도 분산 처리를 적용함으로써 가치 있는 학습 결과를 만들었다. 시스템 전체의 분산 학습 알고리즘은 PSO 알고리즘을 응용하여 자원의 효율성 확보했다[5].

2.5 ImageNet Dataset

ImageNet 이란 어떠한 사진을 보여줬을 때 이 사진이 무엇인지 맞출 수 있는 컴퓨터를 만드는 프로젝트다. ImageNet 에선 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)라고 불리는 이미지 인식 대회를 개최하는데 이는 대용량의 이미지 셋을 주고 이미지 인식 알고리즘의 성능을 평가하는 대회이다. 여기서 사용하는 이미지 셋을 가지고 성능을 검증한다. ImageNet 의 이미지 셋에는 각각 레이블(label)이 붙어있어 분류가 잘 되었는지 평가 할 수 있다[8].

3. 실험

기존 논문에서는 동일한 데이터를 가지고 학습 및 테스트를 진행했기 때문에, 매우 좋은 성능이 도출됐다. 그러나 한정된 데이터이기에 과적합의 우려가 있고, 새로운 데이터가 들어왔을 때 성능을 보장할 수 없다. 또한 많은 양의 데이터로 학습을 진행할 경우 모델의 크기가 크게 증가된다. 이로 인해 발생하는 통신비용을 줄이기 위해 모델의 정확도가 일정 퍼센트 이상 향상되지 않는다면 다운받지 않도록 개선했다.

과적합 문제를 해결하기 위해 검증된 방대한 데이터로 ImageNet 을 선택했다.

ImageNet 의 데이터 중 기존 실험과 유사한 이진 형태의 데이터인 강아지와 고양이데이터를 가지고 분류하는 실험을 했다.

실험에 사용된 머신은 Table 1 과 같다.

Table 1 Machine specification

	CPU	RAM	STORAGE
Name	Intel i5-6400	DDR4 8GB(single)	HDD 1TB

보다 편리한 분산 러닝 시스템 구성을 위해 Oracle Virtual Box 6.0 소프트웨어를 이용해 리눅스 가상머신으로 구성했고, 가상머신의 사양은 코어 1 개를 할당하고 RAM 을 4GB 할당했다. 또한 네트워크 브릿지를 사용해 호스트와 동일한 수준의 IP 주소를 가지고 통신을 진행했다.

CNN[10] 모델을 이용하여 처음부터 모든 과정을 처리하기에는 너무 오랜 시간이 걸린다는 문제가 있어, VGGNet[9]중 Convolutional Layer, Maxpool, 등 16 개의 레이어로 구성되는 VGG16 을 사용하여 이미지의 특징을 추출했다. 이렇게 ImageNet 의 강아지와 고양이 이미지 10400 장을 VGG16 을 통해 전처리했다.

이를 가지고 분류기 모델을 추가적으로 생성해 학습을 진행했다.

분류기의 구조는 Table2 와 같다.

Table 2 Structure of classifier

1 계층	Node:256, Activation Func:Relu, input:8192
2 계층	Dropdout(0.5)
3 계층	Node:1, Activation Func:Sigmoid
컴파일	Optimizer:RMSprop
	Loss:binary_crossentropy

이렇게 전처리된 이미지와 분류기 모델을 가지고 PSO 알고리즘을 적용해 전체적인 시스템을 구성했다.

시스템 동작을 1 회 진행할 때마다, 학습 특성상 발생 할 수 있는 과적합이나, 오차를 줄이기 위하여 학습 데이터와 검증 데이터 및 테스트 데이터를 랜덤으로 섞어줬다. 또한 서버 측에 업로드 되어있는 가장 좋은 정확도를 지닌 모델

을 다운로드 하여 PSO 알고리즘을 통해 모델의 가중치를 수정한다. 그리고 더욱 좋은 정확도를 지닌 모델과 학습에 소모된 시간을 서버에 업로드 했다.

이를 의사코드로 표현하면 Table3 과 같다.

Table 3 pseudo code of system

```
Do
    Do
        Learning
        Calculate fitness Value
        Update local best weight
    While max learning iteration is not attained
    If local best weight > node's best weight then
        Upload local best weight to server
    Upload training time to server
    If local best weight < node's best weight + 0.03 then
        Download node's best weight from server
    Update local weight
While maximum iteration or minimum error criteria is not attained
```

4. 실험결과

1 대의 머신, 5 대의 머신, 10 대의 머신에서 실험을 진행한 결과 1 대의 머신에서 478 초(약 7.9 분)의 학습시간이 소요됐고, 결과는 90.3% 에서 92.8%까지 2.5%의 정확도가 향상됐다. 그리고 5 대의 머신에서 학습시킨 결과 2340 초(1 대당 약 7.8 분)가 소요 되었으며, 정확도는 89.7% 에서 93.2%까지 3.5%의 정확도가 향상됐다. 마지막으로 10 대의 머신에서 실험결과 4908 초(1 대당 약 8.2 분)가 소요됐으며, 정확도는 89.6% 에서 93.4%까지 3.8%의 정확도가 향상되었다. 이렇게 분산 러닝 한 서버의 데이터베이스는 Figure1 과 같다.

num	id	accuracy	time	sec
9	node0	0.896666666667	2019-08-13 15:01:40	92.7331750393
10	node6	0.902916666667	2019-08-13 15:04:02	95.8254711628
11	node1	0.905833333333	2019-08-13 15:04:04	96.3103189468
12	node8	0.915833333333	2019-08-13 15:04:08	95.4465100765
13	node3	0.9125	2019-08-13 15:04:09	91.4945509434
14	node9	0.899166666667	2019-08-13 15:04:09	95.0774929523
15	node7	0.910833333333	2019-08-13 15:04:10	99.4980008602
16	node0	0.920833333333	2019-08-13 15:05:29	93.0028569698
17	node3	0.930416666667	2019-08-13 15:05:48	95.448143959
18	node3	0.934166666667	2019-08-13 15:09:35	91.3642060757

Figure 1 distributed result of 10 machines

Figure 1 의 num 은 업로드된 순서, id 는 분산 시스템에서 성능이 가장 좋았던 노드의 이름, accuracy 는 그 해당 노드가 업로드한 정확도, time 은 해당 기록이 업로드된 시점, sec 는 해당 노드가 정확도를 만들기 위해 학습한 시간을 나타낸다.

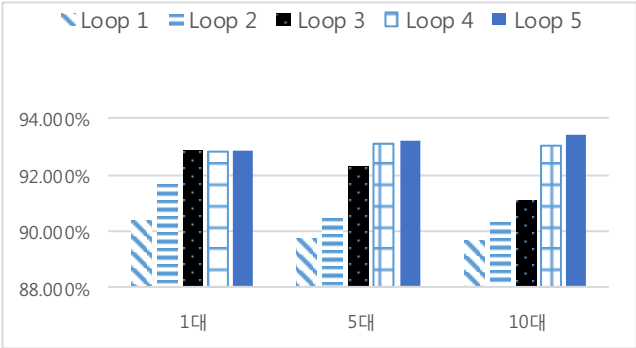

Figure 2 performance chart

Figure 2 는 Figure1 의 결과를 차트로 정리한 것이며, x 축은 분산 처리에 사용된 머신의 개수이며, y 축은 정확도를 나타낸다. 막대 차트의 패턴은 각각 Loop Count 를 나타낸다.

5. 결론

ImageNet 데이터를 이용하여 PSO 알고리즘 분산 러닝 시스템을 검증한 결과, 분산 시스템의 머신이 많아질수록 1 대에서는 92.833%였지만, 10 대의 머신에서는 93.417%로 머신의 개수가 늘어남으로써 약 0.6% 정확도 향상이 이뤄졌다.

실험결과, 분산 러닝 시스템으로 학습을 진행할 경우, 더욱 다양한 변수가 발생할 수 있어 단일에서의 학습에 비해 더욱 좋거나, 나쁜 결과가 도출될 수 있다. 또한 코드 내부에 넣은 트레이닝 셋과 테스트 셋, 검증 셋을 랜덤으로 섞어주는 과정으로 인하여 과적합 (Overfitting) 을 최소화할 수 있다.

본 논문은 검증된 ImageNet 의 방대한 데이터를 사용했지만 데이터 증식을 사용하지 못했다. 사용하지 못한 이유는 머신들의 낮은 사양과, 서버 시스템의부하, 시간적으로 한계가 있다. 그래서 추후 시스템을 개선 및 보완하여 데이터 증식을 사용한 실험을 진행하고자 한다.

참고문헌

[1] MIT, 2013 년 10 대 혁신기술 선정 : http://www.donga.com/news/article/all/20130426/54713529/1

[2] 자율주행 자동차, '딥러닝'으로 시동건다. 삼성뉴스룸 https://news.samsung.com/kr/?p=368515

[3] 구글 신경망 번역 시스템 GNMT.https://ai.google/research/pubs/pub45610

[4] 권대책, 강보영. (2017). 합성곱 신경망 사용을 위한 CPU 와 GPU 성능 분석. 한국정보기술학회논문지, 15(8), 11-18.

[5] 조인령 "PSO 알고리즘을 이용한 분산 딥러닝 시스템" 한국정보처리학회 2017 추계학술대회 논문

[6] 김영환 "효과적인 동기화 기법을 활용한 분산 병

렬 딥러닝 구조" 한국외국어대학교 석사학위청구 논문
[7] 조태호 저 “모두의 딥러닝”
[8] Olga Russakovsky 저 “ImageNet Large Scale Visual Recognition Challenge”
[9] Wei Yu 저 ”Visualizing and Comparing AlexNet and VGG using Deconvolutional Layers”
[10] Quan Zhang 저 “Convolutional Neural Network”
[11] Gerhard Venter 저 “Particle Swarm Optimization”
[12] Pima-indians-diabetes-database, https://www.kaggle.com/uciml/pima-indians-diabetes-database
[13] 박평재. “PSO 최적화 기법을 이용한 Ethylene Oxide Plant 배치에 관한 연구”, 『한국안전학회지』, 2015.